学校编码：10384　　分类号_密级_

学号：15420131152031　　UDC_

硕士　学　位　论　文

稀疏聚类方法在中国股票市场中的应用研究

The Application of Sparse Clustering Algorithm in Chinese Stock Market

张悦涵

指导教师姓名：谢邦昌　教授

专　业　名　称：统　计　学

论文提交日期：2016 年 2 月

论文答辩时间：2016 年 4 月

学位授予日期：2016 年 6 月

答辩委员会主席：_____

评　　阅　　人：_____

2016 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。(请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。)

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人：

年　月　日

摘要

随着中国人日常生活水平的提高，人们已不再满足于拿固定的工资，越来越多的公民选择将自己手中的资金投资到股票市场，希望能够钱生钱。但并非所有股民都是金融专业人员，如何进行股票种类的选择与资金的分配已经成为越来越多的中国人所希望知道的问题。选择股票的两个基本原则就是低风险、高收益，那么如何在上万只股票中选择最适合自己的股票，如何缩小选择范围，如在某个行业内选择股票是本文所希望解决的问题。

众所周知，股票数据是动态的，随时间的推移而不断变化，因此传统分析股票的方法为时间序列分析法，基于随机过程理论和数理统计学方法，找出股票数据的内在规律，以解决实际应用的问题。时间序列分析最大的优点是实际数据的时间序列能够展示对象在一定时期内的发展变化趋势与规律，因而可以从时间序列中找出变量变化的特征、趋势以及发展规律，借此对变量未来的变化进行有效的预测。但时间序列分析法突出了时间因素在预测中的作用，暂不考虑其他外界因素，因而预测误差较大，当数据的其他变量发生较大的变化时，时间序列分析法就没有那么有效了。因此，时间序列分析往往针对短期的数据，短期其他因素发生变化的概率较小。

本文针对以上的问题，将股票数据随时间推移的特点与其他指标，如开盘价、收盘价、交易量、涨跌幅等进行结合，将 T 个时间点的 m 个指标看作是 T*m 个变量，这样不仅解决了时间序列分析法的局限性，还可以分析长期股票的变化，对股票未来的发展进行更准确的预测。但又出现了新的问题，经过转换后的股票数据变成了高维甚至超高维的数据，并且不一定所有时间点都会对股票的变化形势产生影响，此时，通过变量选择将数据进行降维是必要的。本文将稀疏聚类方法应用于股票数据中，对变量赋予不同的权重，然后对权重添加惩罚因子，如 Lasso 惩罚因子，求解权重的最优值，根据权重的大小分析变量对聚类结果的影响，当权重约等于零时，这个变量对聚类结果没有影响，可以剔除。相对于主成分分析法，稀疏聚类真正做到了变量筛选，而非将原始变量进行线性组合形成新的变量。同时，稀疏聚类还可以将股票数据进行分类，比如，将股票根据涨跌幅或者交易量分为不同的类，使股民缩小选择范围。相比于传统的一般聚类方法，

本文通过模拟研究证实了稀疏聚类在处理高维数据方面具有更好的效果。

本文第一章是绪论部分，介绍了研究的背景与目的，并说明了本文的研究结构。

第二章是文献综述，到目前为止，已经有不少文章做了关于变量选择方面的研究，在这一章本文对此进行了回顾，并指出了其不足之处；除此以外，股票数据也不是新的研究领域，本文也对前人的研究做了汇总说明。

第三章是相关理论介绍，首先介绍了数据挖掘理论，并阐述了数据挖掘理论与统计学的结合；其次是聚类理论，并对聚类中一些概念的定义作了说明；再次是最常见的聚类方法之一，K-means 聚类的介绍；接着阐述了层次聚类的算法原理、步骤等；最后在引出了稀疏聚类的概念，包括稀疏 k-means 聚类和稀疏层次聚类，并详细介绍了其算法思想、原理以及步骤。除此以外，在此章节还对本文所用软件进行了说明。

第四章是数据模拟研究，为了证明稀疏聚类的有效性，本文进行了四个部分的模拟研究。首先是证明稀疏聚类在变量选择方面的有效性，通过生成随机数据，将稀疏聚类所画出的 Gap 统计量图和权重散点图与真实情况作对比；其次是分析不同数据结构对稀疏聚类效果的影响，这部分模拟研究又分成三部分，分别是观测量和变量数、数据均值与变量数以及数据均值与方差对稀疏聚类的影响；接着是稀疏 k-means 聚类对分类的效果，将其与一般 K-means 聚类进行比较，通过分类错误率的标准，判断出对于不同维度的数据，稀疏 K-means 聚类与一般 K-means 聚类各有其优势；最后对生成的随机数据分别应用稀疏层次聚类和一般层次聚类，通过树图对其效果进行判断。

第五部分是实证研究，通过选取万得咨询中上证 A 股的数据，首先进行数据处理，进行了数据筛选与标准话等处理；其次选取了 100 股作为样本，在 1000 个变量的基础上对其进行聚类，运用稀疏聚类方法；首先运用稀疏 k-means 聚类方法，筛选出了 125 个交易日的成交量作为对结果有影响的变量，即聚类结果是基于成交量进行的；其次对实证数据运用稀疏层次聚类，在变量筛选的同时，画出了聚类结果的树图，将数据分为五类；最后为验证稀疏聚类的效果，对数据分别应用了一般聚类方法。在验证方面，通过对真实数据画折线图与聚类结果进行对比，得出聚类结果的准确性的结论；同时，通过 Dunn 指数判定了四种聚类方

法效果的好坏。

第六部分是结论与建议，首先对研究过程进行了总结；其次对成交量最高的第五类的企业进行了分析，在此基础上对股民提出了建议。

最后一部分是本文的创新点与不足，并提出了后续研究建议。

关键字：变量选择；稀疏聚类；股票市场

Abstract

With the development of quality of lives among Chinese people, more and more citizens are no longer satisfied with the mere salaries, many of them choose to turn to the stock market to make their money derive larger amount of money. However, not everyone is professional in finance, most of them need advice on how to choose stocks and how to allocate their capital. Two basic principles of choosing stocks is high value, low risk. It is not so easy to choose stocks with high value and low risk among thousands of types. In this article, we hope to solve the problem of diminish the hope of stock candidates to help investors make better choices.

It is widely known that stock data is dynamic, it changes with time. Thus analysis of time series is the most common method to analyze and predict the trend of stocks. The analysis of time series is a statistic method based on the theory of random processes and mathematical analysis. It helps find the inner regulars of stock data and solve the practical problems. One of the greatest advantages of analysis of time series is that the time series of real data can show how they perform in a certain period of time, through this, we can find the character, trend and development regulars of features. However, the analysis of time series highlights the role of time and chooses to ignore the other factors among its analysis, which leads to high error rate of prediction. Especially when predicting changing trend in a long period, as the possibility of alteration of other variables is quite large, the analysis of time series is certainly not so effective.

In order to overcome the problems above, in this article, time points T is combined with the indices of stocks, such as opening price, highest price, exchange money, performance of price, etc. through this way, we can regard the m indices of stock data in T period as T*m features. It not only solve the problems of limitations of analysis of time series, but as can solve the problems in long periods. Whereas there comes another problem, the data becomes high or even super high dimensional data, and we cannot guarantee that every features is influential for the classification of stock data.

According to this, it is necessary to reduce dimensions of the data. In this article, sparse k-means method is applied in the stock data. In sparse k-means method, feature is weighted, and the weight is bound to penalty, such as lasso penalty. By comparing the weights of features, one can choose which feature has greater influence on the result of clustering and which one can be omitted as it has nothing to do with the result. Compared to principle component analysis, sparse k-means can result in real screening of features, rather than combine the features with linear methods to produce new features. Meanwhile, data is clustered by sparse k-means method. For instance, cluster the stock data based on performance of prices or exchange quantity, and help investors diminish their choosing scope. Sparse k-means method has proved to be more effective in dealing with high-dimensional data in this article.

The first chapter of the article is introduction. Study background and directions are introduced in this chapter, it also state the structure of the article.

The next chapter is literature review. So far, a number of authors have noticed the importance of feature selection and do a lot of work in this field. This chapter makes an overview of them and points out the drawbacks of these methods. Apart from this, stock data is not a new study field either. A summary review of previous studies is also shown in this chapter.

The third chapter is the about theory. First is the theory of data mining, and state the relationship between data mining and statistics. Second is the theory of cluster, as well as some concepts in it; then we introduce the theory of one of the most common methods of clustering, k-means clustering; last is the concept of sparse k-means, we describe the thought, principle, and steps of sparse k-means method in detail. Software uses are also introduced in this chapter.

Simulation study is in the fourth chapter. In this article, three parts of simulations are studied to prove the efficiency of sparse k-means method. First of all, by generating random data, apply sparse k-means on it, and compare the results with the real number of features, we proved that the sparse k-means method is effective in feature selection; second, to solve the problem of how data structure affect the efficiency of sparse k-means, we did three simulations, with different numbers of observations and

numbers of features, different mean values and numbers of features, and different mean values and variation values respective; last is comparing the result of sparse k-means and regular k-means, using the criterion of classification error rate (CER), to prove that for different dimensions of data, sparse k-means and regular k-means has their own advantages.

Chapter five is experimental structure. In this part, we download data from WIND, and choose the stock data of Shanghai A stock market. The first step is data pretreatment, we choose the data that has been on Shanghai A stock market through ten years, and standardize the data. Then we choose 200 data as the sample and the number of features is 960. After all the dealing, we first apply sparse k-means method in the data and select the influential features. To guarantee the efficiency of sparse k-means in feature selection, we also perform principle component analysis on the data, and compare the two results. Then we apply regular k-means method on the data and compare the two clustering results, using two kinds of criterions, CER and Dunn index.

Chapter six is conclusion and suggestion. We present some of the clustering result in this chapter, and according to this, we suggest investors to invest stocks in the career of engineering. Then we present the news in the latest couple of years and prove the accuracy of our result. Last, we make some conclusions about the article

The last chapter is innovation points and deficiency of the article, future studies are also suggested in this part.

Key words: Feature selection Sparse Clustering Stock Market

目 录

Content

第 1 章 绪论

1.1 选题背景与目的

改革建设以来，我国经济市场一直在高速发展，市场制度也在不断完善，越来越多的公民已经不再满足于拿死工资，而是希望进行适量的投资理财，达到钱生钱的目的。股票市场近年来炙手可热，对于投资者来说，选择股票的两个基本原则是高收益、低风险，因此，通过对股票的分析和对市场的解读是股民投资股票的基本功。近几年来,证券市场不断扩容、上市公司数量急剧增加，面对成千上万的股票，投资者需要对其进行分类来缩小自己的选择范围，从而做出更为理智的选择。聚类分析是一种常见的分组方式。它可以将“相似程度”高的股票分为一组，将“相似程度”低的股票分为不同的组，帮助投资者更为直观的看到股票的特征和发展趋势，从而更加理智的根据自己的实际需要对股票进行选择与资金分配。

通常所说的股票数据一般考虑两种类型：股票行情数据和客户交易数据。分析股票的变动趋势以及行情的发展前景主要运用股票行情数据，包括股票的开盘价等各种价格指标、交易量、涨跌幅等。通过股票价格，可以分析出其变动的内在规律，进而分析股票价格的长期走势。股票价格是一组按时间顺序排列的数据,称其为时间序列数据。这类数据反映了股票价格随时间变动的状态或者程度，时间序列数据可以细分为季度数据、月度数据等，具有周期性、循环性、趋势性、变动性等特点。目前，对于时间序列数据已经发展出一系列的方法来对其进行分析、挖掘，例如随着计算机技术的发展以及数据挖掘技术的成熟，目前已经有了系统的针对时间序列数据的数据挖掘方法。时间序列数据的作用主要是根据历史数据对数据未来的变化趋势做预测，以时间序列数据为目标数据，通过对其进行统计分析，从数据中找出内在的变化趋势与规律，进而对指标未来的变化做预测。在数据挖掘的众多方法中，聚类是一种重要手段,它是把一组物理的或抽象的对象按照相似性进行归类,也称为“无指导分类”。聚类的分类标准为将类内距离最小化、类间距离最大化，距离的定义按照具体情况决定，最常用的是欧几里得距离。同时，由于聚类无需事先定义好类别和类数，相对于分类方法约束条件少很多，因此已经成为越来越常见的工具。目前,有三类聚类方法适合对股票价格时

间序列数据进行聚类：基于原始数据的方法、基于模型的方法和基于特征提取的方法。这些方法都是针对单支股票时间序列而提出的,用来对一个时间序列的各个子序列进行聚类。首先对时间序列进行分割,然后对分割后得到的子序列集进行聚类。这种方法对单只股票的波动性、预测这只股票的价格变动趋势并帮助投资人做出正确的决定是有用的。

当今社会，随着人们风险意识的提高，“鸡蛋不能放进同一个篮子”已经成为大家公认的规避风险定理。股民们会选择多只股票进行投资，对于投资公司来说，更是会选择大量的股票。在这种情况下，对股票时间序列而数据进行聚类分析的方法不仅会导致工作量的大量增加，还会因为对单只股票进行分析忽略了股票间变化的相关性，而导致投资者不能确定投资金额的分配。除此以外，时间序列分析法强调时间的作用，却并未对其他因素予以适当的考虑，即此方法是以其他因素固定不变为假设条件的，在长期，这个假设很难成立，因而会导致很大的预测误差。

本文将从另外的角度看待股票行情数据，将每个时间点每只股票的行情数据看作不同的变量，例如，对 n 只股票 T 个时间点的数据进行分析，每只股票有 q 个维度，那么数据矩阵由（n*T, q）变为（n，T*q）。以此为前提，对 n 只股票进行聚类，选出高收益、低风险的股票组合进行投资，还能同时确定投资金额的分配。

然而，并不能直接将传统的聚类方法应用于此时的数据，因为其具有数据量大、维度高的特点，并且不一定所有的时间点都对股票的走势产生决定性的影响，如果用全部变量对数据进行聚类，则可能导致重要类别的丢失。本文第一次采用稀疏聚类的方法分析股票数据，在此方法中，对变量赋予权重，并对权重添加 lasso 惩罚因子，压缩权重，进而根据其的大小对变量的重要性进行排序。在此框架的基础上，用稀疏聚类的方法对股票数据进行聚类，并根据单一的标准来控制变量选择和聚类的结果。

1.2 研究方法与步骤

围绕研究目的，整个文章可以分为两部分的研究：理论研究和实证研究。

理论研究从数据挖掘、聚类、稀疏聚类由大到小的概念介绍，尤其对稀疏聚类，包括稀疏 K-means 聚类和稀疏层次聚类两种方法的概念、原理、思想、步

骤等做了详细的阐述。为证明稀疏聚类的有效性，本文还进行了模拟研究。

实证研究选取 2015 年 7 月 1 日到 2015 年 12 月 31 日半年来的上证 A 股股票数据，以天为单位，股票指标选择开盘价、最高价、最低价、收盘价、成交量、涨跌和涨跌幅这些常用指标。在 1076 支股票数据中，抽取 100 支股票作为样本，同时，半年的交易日有 125 天，因此共有 125*8=1000 个变量。在实证分析中，首先对其进行稀疏 K-means 聚类和稀疏层次聚类，进行变量选择，根据变量权重图，可以明显看出成交量对聚类结果有影响，并且每天的影响都不一样，前 50 天左右的成交量的影响要大于后 75 天；其次，将聚类结果用表格的形式表示出来，与同一类的股票的成交量折线图作对比，证明分类的准确性；接着为了证明稀疏聚类方法的效果优于一般聚类方法，又对数据进行了一般 k-means 聚类和一般层次聚类；最后选用了 Dunn 指数来判定稀疏聚类与一般聚类的效果。

1.3 文章结构

本文共分为四个部分，第一部分为绪论和文献综述，引出文章主题，稀疏聚类方法，并指出将此方法应用到股票数据的可行性。第二部分为理论部分，包括第三章相关理论介绍和第四章数据模拟研究，分别从理论和实践的角度对稀疏 K-means 聚类的方法进行了详细的阐述。第三部分为实证研究部分，通过对股票数据进行稀疏 K-means 聚类，并将其结果与其他方法进行比较，验证可行性，并证明稀疏 K-means 聚类方法的优势。第四部分为结论部分，通过实证分析的结论，对投资者做出建议，同时验证结论的正确性，并对文章作了总结以及对未来的研究方向做出了展望。